DOCUMENT RESUME

ED 278 707 TM 870 134

AUTHOR Fish, Larry
TITLE The Importance of Invariance Procedures as against Tests of Statistical Significance.
PUB DATE Nov 86
NOTE 25p.; Paper presented at the Annual Meeting of the Mid-South Educational Research Association (Memphis, TN, November 20, 1986).
PUB TYPE Speeches/Conference Papers (150) -- Reports - Evaluative/Feasibility (142)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Correlation; Hypothesis Testing; *Multiple Regression Analysis; Multivariate Analysis; Predictor Variables; Research Methodology; *Social Science Research; *Statistical Significance
IDENTIFIERS *Cross Validation; Null Hypothesis; Research Replication

ABSTRACT

A growing controversy surrounds the strict interpretation of statistical significance tests in social research. Statistical significance tests fail in particular to provide estimates for the stability of research results. Methods that do provide such estimates are known as invariance or cross-validation procedures. Invariance analysis is largely an untested science which is applied to determine how stable the statistical results are likely to be across different samples. It can be employed with any parametric procedure. The details of invariance analysis vary according to the analytic technique employed. Cross-validation procedures appropriate for multiple regression and its multivariate extension, canonical correlation analysis, are discussed in this paper, and a concrete example is presented. (Author/JAZ)

**********
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
**********

ED278707

# THE IMPORTANCE OF INVARIANCE PROCEDURES AS AGAINST TESTS OF STATISTICAL SIGNIFICANCE

Larry Fish

University of New Orleans

Paper presented at the annual meeting of the Mid-South Educational Research Association, Memphis, Nov. 20, 1986.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

L. Fish

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

X This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

TM 870 134

BEST COPY AVAILABLE

ABSTRACT

A growing controversy surrounds the strict interpretation of statistical significance tests in social research. Statistical significance tests fail in particular to provide estimates for the stability of research results. Methods that do provide such estimates are known as invariance or cross-validation procedures, and they can be applied in most analyses. Cross-validation procedures appropriate for multiple regression and its multivariate extension, canonical correlation analysis, are discussed in this paper, and a concrete example is presented.

> If we take in our hands any volume of school metaphysics, for instance, let us ask, 'Does it contain any abstract reasoning concerning quantity or number?' No. 'Does it contain any experimental reasoning concerning matter of fact and existence?' No. Commit it then to the flames, for it can contain nothing but sophistry and illusion.
>
> ---David Hume (quoted in Will Durant, The Story of Philosophy.)

Statistical significance testing is the "experimental reasoning" of choice among most researchers today, and while its absence in an empirical study may no longer be cause for commitment to flames, it may result in notices of rejection from publishers or from dissertation committees. Nearly 30 years ago, however, Selvin (1957) publicly questioned the value of statistical significance testing as an inferential tool in social research. Selvin's article initiated a controversy which continues to this day, with increasingly formidable artillery ranged on the side of significance testing's opponents.

The philosophy of statistical significance testing assumes an abstract simplification of the reality in which social scientists are interested. In a universe of human behavior shaped by complex relationships among large numbers of variables, the statistical significance test can only provide a binary solution--that is, a simple "yes or no" answer--to a single question of relatively little inherent interest--is the null hypothesis to be rejected?

The logic of statistical significance testing is at first

compelling, for it is based on the perfectly reasonable assumption that the larger two random samples are, the closer should be their means an any measure of interest, provided that the samples are from the same population. However, the mathematical dependence of statistical significnce upon sample size can make even negligible research results appear "important." Carver (1978) observed that "a mean difference that is small and not significant from a research standpoint can be statistically significant just becuse enough subjects were used in the experiment to make the result statistically rare under the null hypothesis" (p. 388).

The typical null hypothesis, which postulates the absence of "variance explained," is usually of little inherent interest. Furthermore, rejecting a null hypothesis is generally done on the basis of criteria--for example, the 5% significance level--which, however reasonable they may be, are nonetheless arbitrary. Too much light focused on the "significance" of a null hypothesis can leave the most meaningful implications of an experiment in total darkness. Lykken (1968) warned that "[Finding statistical significance] is never a sufficient condition for concluding that a theory has been corroborated, that a useful empirical fact has been established with reasonable confidence--or that an experimental report ought to be published" (p. ).

Another problem confronted in statistical significance testing is that, as Cotton (1967, p. 57) pointed out, the null

hypothesis in social research is never wrong: rarely, if ever, will two variables share a correlation coefficient exactly equal to zero. Cotton argued that "accepting [a null hypothesis] merely expresses a belief that the averge difference is near zero" (p. 57), but as alredy noted, even average differences that are "near zero" can be forced to assume unwarranted significance when samples are large enough. Furthermore, apart from the fact that it usually cannot be wrong, the null hypothesis in any given experiment is but one of an infinite number of possible research hypotheses, and rarely is it the most illuminating one.

The most interesting research results are those which, however significant statistically, can be generalized from a sample to a larger population. Small relationships that are consistent over samples are of potentially greater theoretical interest than are pronounced relationships that can be obtained for only one or two samples. Of course, the ideal research result would be large relationships that are consistent over samples. However, Carver (1978) has argued that statistical significance is not an index of reproducibility: statistical significance at the level P does not necessarily imply a probability of (1 - P) that another researcher following the same procedures will obtain the same results. The confounding of statistical significance and reproducibility, argued Carver, is one of three major prevailing misconceptions about statistical significance testing, the other two being that a statistical significance level represents the probability that results were obtained by chance, and that it represents the probability that

the null hypothesis is true.

Obviously, the only genuine way to establish the replicability of research results is actually to repeat the study on as many samples as possible. Unfortunately, this is rarely practical. However, the researcher can still obtain an estimate of the stability of his results across samples by employing so-called invariance or cross-validation procedures. These procedures are the subject of the discusion that follows.

The logic of invariance analysis was summarized by Fish (1986) as follows:

> [Invariance procedures] attempt to determine how stable the statistical results are likely to be across different samples. In the typical invariance procedure an analysis is performed separately on each of two subgroups into which the study sample has been divided, and the results are compared. When the results of an analysis are not comparable--i.e., not invariant--serious doubts about the generalizability of the results are in order. (pp. 65-66)

Apart from its value to theory building, a successful invariance analysis will create confidence that analytic results can be employed for practical ends. "Double cross-validation," argue Kerlinger and Pedhazur (1973), "is strongly recommended as the most rigorous approach to the validation of results from regression analysis in a predictive framework" (p. 284).

Invariance procedures can be employed profitbly with any parametric procedure. The details of invariance analysis vary according to the analytic technique employed, and because invariance analysis is still a relatively young technique, there is ample scope for imaginitive applications. The remainder of this paper will focus on standard cross-validation procedures that can be employed with multiple regression and with its multivariate generalization, canonical correlation analysis. A concrete example will be discussed.

## Invariance procedures for multiple regression

Recall that when there is one dependent variable y and two ormore independent variables x(i),multiple regression analysis computes for each case (person) a composite score y' which is equal to a linear combination of that case's values on the independent variables, as follows:

$$\sum_i \beta_i x_i = y' \qquad (1)$$

For the entire study sample, the squared correlation coefficient between the composite scores and the actual values of the dependent variable is a measure of effect size--that is, the proportion of variance of the dependent variable that is shared with the independent variables.

The invariance procedure for multiple regression consists of the following steps:

1) The original sample is randomly divided into two invariance groups. Ideally, these two groups are of unequal size so as to obviate the objection that a satisfactory measure of invariance is dependent upon a particular sample size.

2) Within each of the two groups, all variables are separately standardized into z-scores and independent regression equations are computed. For each case, an invariance composite score (y'(1,1) for cases in group 1, y'(2,2) for cases in group 2--the meaning of the double subscript will become clear shortly) is computed from the appropriate equation.

Group 1: Group 2:

$$\sum_j \beta_{1j} x_{1j} = y'_{11} \qquad \sum_j \beta_{2j} x_{2j} = y'_{22} \qquad (2, 3)$$

3) We now proceed to establish the invariance of the multiple regression equation computed for invariance group 1. We have alredy computed a set of composite scores y'(1,1) for the cases in this group. We now compute a second set of composite scores for each case, y'(1,2), using the beta-weights computed for invariance group 2. This is the key step of the invariance procedure.

$$\sum_j \beta_{2j} x_{1j} = y'_{12} \qquad (4)$$

The subscript of this new composite score refers to the fact that group 1 data were applied to group 2 regression coefficients. Throughout this paper, the subscript ij appearing below any

composite score or correlation coefficient means that group i data were applied to group j weights. (Recall that earlier, $y'(1,1)$ referred to composite scores computed from group 1 data dubstituted into the group 1 regression equation.) However, when the subscript ij appears below $\beta$ or x, it refers to group i, independent variable number j.

4) For group 1 we may now compute two multiple regression coefficients: the group's own "genuine" coefficient R(1,1) between the set of y and the set of $y'(1,1)$, and an <u>invariance correlation coefficient</u> R(1,2) between the set of y and the set of $y'(1,2)$. R(1,2) cannot exceed R(1,1) because the latter is the mathematical optimum for group 1, but ideally the two coefficients will be very close. The difference between the squares of these two coefficients, $R(1,1)^2 - R(1,2)^2$ (recall that only squared correlation coefficients can be meaningfully compared) is an invariance estimate. The closer this value is to zero, the more stable the regression results may be assumed to be across samples. The correlation coefficient between the sets of composite scores $y'(1,1)$ and $y'(1,2)$ may also be taken as an invariance estimate.

Naturally the procedures outlined above can be repeated with the roles of groups 1 and 2 reversed--that is, group 2 data can be substituted into the group 1 regression equation, and an invariance estimate computed for the group 2 regression equation. This would complete the invariance procedure known as "double cross-validation."

Some comments are in order here. For one thing, it may be objected that the procedure decribed above establishes the invariance of the two group regression equations, not of the omnibus equation (1) that is presumably of primary interest. This objection has some merit, and reflects the status of invariance analysis as an imperfect substitute for genuine replication. However, both Mosier (1951) and Kerlinger and Pedhazur (1973) have argued that when the results of double cross-validation are satisfactory the omnibus regression equation may be confidently employed for predictive purposes. Presumably it may also then be used for theoretical purposes as well.

It may also be argued that no specific criteria were offered in the above discussion for evaluating the invariance estimates. This omission was deliberate, for no such criteria yet exist. As mentioned above, invariance analysis is a relatively young procedure, and many avenues remain to be explored. Thompson (1986), for example, has derived test criteria for invariance estimates computed for factor analysis. However, in some respects it is illogical to test the statistical significance of results that in some senses are meant to replace significance testing.

One other item of useful information that can be derived from invariance analysis concerns multiple R. As mentioned above, multiple R is a mathematical optimum for a given sample, and as such it is a biased estimate that capitalizes on what Mosier called the "idiosyncracies" of the sample. Naturally a more

dependable estimate of multiple R for the target population would be desirable. Though no precise means yet exist for computing this more dependable estimate from invariance data, Mosier suggested that the mean of an invariance group's actual squared multiple R and the square of its invariance R might be taken as a provisional estimate.

Invariance procedures for canonical correlation analysis

Before discussing invariance procedures in canonical correlation analysis, it would be useful to recall that canonical correlation analysis is a multivariate generalization of multiple regression. Indeed, as argued in Thompson (1984), all parametric techniques are special cases of canonical correlation. Canonical correlation is the appropriate analytic technique when each variable set--predictor (independent) and criterion (dependent) variables--hastwoormore elements.

As in multiple regression, canonical correlation analysis computes for each case a predictor composite value p' equal to a linear combination of the independent variables, x(i). Analogously, it computes a criterion composite value q' equal to a linear combination of the dependent or criterion variables, y(i). Naturally, as in multiple regression, the same function coefficients are used for all the cases in the sample.

Predictor composite value: Criterion composite value:

$$\sum_i a_i x_i = p' \qquad \sum_i b_i y_i = q' \qquad (5, 6)$$

The correlation coefficient Rc between the set of predictor composites p′ and the set of criterion composites q′ is the canonical correlation coefficient. According to Thompson (1984), "a squared canonical correlation coefficient indicates the proportion of variance that the two composites derived from the two variable sets linearly share" (p. 14). It should be clear that this squared canonical correlation coefficient is the analog of multiple R squared in regression analysis.

The two linear equations (equations 5 and 6) which give, respectively, the predictor and criterion composite scores are known together as a canonical function. The linear coefficients of a canonical function are derived so as to maximize the shared variance between the two composites for any function. It should be noted that more than one canonical function may be derived in an analysis, the number of such functions being equal to the number of variables in the smaller of the two variable sets. Each canonical function derived after the first maximizes the explained portion of variance not yet accounted for by any of the previously derived functions. The reader interested in pursuing the logic of canonical correlation analysis further should consult Thompson, 1984.

The logic of invariance analysis in canonical correlation is essentially the same as for multiple regression, discussed above. Once again, the sample is randomly divided into two invariance groups of unequal size. Within each of these two groups, the values of all variables are converted to z-score form and

independent canonical correlation analyses are conducted. Because several canonical functions may be derived, it may be necessary to repeat the invariance procedure described below on as many of those functions as are considered to be statistically or educationally significant.

Let us assume, then, that within each invariance group we have derived a canonical function, as follows:

Group 1:

Predictor composite:

$$\sum_{i} a_{1i}x_{1i} = p'_{11}$$

Criterion composite:

$$\sum_{i} b_{1i}y_{1i} = q'_{11} \quad (7, 8)$$

Group 2:

Predictor composite:

$$\sum_{i} a_{2i}x_{2i} = p'_{22}$$

Criterion composite:

$$\sum_{i} b_{2i}y_{2i} = q'_{22} \quad (9, 10)$$

In the above equations, a(k,i) and b(k,i) represent respectively the standardized predictor and the standardized criterion function coefficients derived for invariance group k and variable i. In symbols representing composite scores and canonical correlation coefficients (that is, the letters p′, q′ and Rc), the double subscript ij means that the coefficient in question

was derived from group i data using group j coefficients. This is the same notational format that was used earlier in the discussion of invariance for multiple regression.

We shall now investigate the invariance of the group 1 canonical function, always bearing in mind that the same procedure should be applied afterwards to the group 2 function as well. Thus within group 1, two more sets of composite scores, $p'(1,2)$ and $q'(1,2)$, will be computed using group 1 data but group 2 function coefficients, as follows:

$$\sum_j a_{2j}x_{2j} = p'_{12} \qquad \sum_j b_{2j}x_{2j} = q'_{12} \qquad (11, 12)$$

A "new" canonical correlation coefficient, Rc(1,2), is computed for the two new sets of composite scores, the set of $p'(1,2)$ and the set of $q'(1,2)$. Because Rc(1,1) is the mathematical optimum for group 1, it must be at least as large in absolute value as Rc(1,2). The difference between the squares of Rc(1,1) and Rc(1,2) is an invariance estimate for the group 1 canonical function. As in the case of multiple regression, this estimate will have a "best case" value of zero and a "worst case" value of 1.

The reader should note that the invariance procedure for canonical correlation analysis is analogous in almost every detail to the procedure discussed above for multiple regression; only the vocabulary is different. In fact, if the set of criterion variables in canonical correlation analyiss contains

only one member y, then canonical correlation analysis reduces to multiple regression. The original dependent variable takes the place of the criterion composite, and multiple R takes of place of the canonical correlation coefficient.

An illustrative example

This section illustrates the computation of an invariance estimate for canonical correlation analysis. Table 1 presents the hypothetical data set that will be used. This data set is small enough so that the reader, if interested, may follow the discussion with pencil and paper. Each variable set, predictors and criterion variables, contains two variables, and canonical correlation analysis will therefore yield two functions. The invariance of only the first function will be discussed below; the interested reader may wish to apply invariance procedures to the second function an an exercise.

The first step of the invariance procedure is to divide the sample into two invariance groups. For convenience, the first five cases of the hypothetical data set have been placed into group 1, and the second five into group 2, though ideally two invariance groups should be randomly assigned and of unequal size. Table 2 presents the values of the variables ater being converted to z-score form within each group.

The next step of the procedure is to compute separate canonical correlation analyses for each of the two invariance groups. This can be done effectively only with a computer.

A complete canonical correlation analysis yields a considerable amount of data; of these data, only the standardized function coefficients are immediately relevant to our invariance procedure. These coefficients are presented in Table 3.

Table 4 presents the data that will be used to compute an invariance estimate for the first canonical function in group 1. The first two columns of Table 4 represent respectively, for each invariance group, the predictor and criterion composite scores as computed from equations 7 - 10. The third and fourth columns present the invariance composite scores which, within each group, were computed from the other group's equations (equations 11 and 12 for group 1). The following four equations illustrate how these values were computed for case 1, group 1.

Predictor composite:

$$( 0.971)(-0.878) + ( 1.146)( 1.321) = 0.661$$

Criterion composite:

$$( 1.373)( 0.309) + ( 0.751)( 0.288) = 0.641$$

Invariance predictor composite:

$$( 0.042)(-0.878) + ( 0.989)( 1.321) = 1.269$$

Invariance criterion composite:

$$( 1.174)( 0.309) + (-0.249)( 0.288) = 0.291$$

In the above equations, the data came from Table 2, and the

coefficients from Table 3.

We now have all the data necessary to compute the following four correlation coefficients:

1) Rc(1,1): the "actual" canonical correlation coefficient for group 1; the correlation coefficient between group 1 predictor and criterion composite scores (Table 4, columns 1 and 2).

2) Rc(1,2): the invariance correlation coefficient for group 1; the correlation coefficient between group 1 invariance predictorand invariance criterion composite scores. (Table 4, columns 3 and 4.)

3) Rc(2,2): the "actual" canonical correlation coefficient for group 2; the correlation coefficient between group 2 predictor and criterion composite scores (Table 4, columns 1 and 2).

4) Rc(2,1): the invariance correlation coefficient for group 2; the correlation coefficient between the invariance predictor and the invariance criterion composite scores (Table 4, columns 3 and 4).

Table 5 presents the squares of these four correlation coefficients in the following format:

| | Source of function coefficients: | |
|---|---|---|
| | Group 1 | Group 2 |
| Source of data: | | |
| Group 1 | $Rc(1,1)^2$ | $Rc(1,2)^2$ |
| Group 2 | $Rc(2,1)^2$ | $Rc(2,2)^2$ |

The difference between the two entries in the first row is an invariance estimate for function 1, group 1, while the difference between the two entries in the second row is an invariance estimate for function 1, group 2. Considering the extremely small size of the data set, these invariance estimates--0.424 and 0.357 for groups 1 and 2 respectively--are not too bad. In a real research situation with a much larger sample, one would naturally hope for smaller estimates.

Before closing it is worth pointing out again that invariance analysis is still a young and largely untested science, and the interpretation if invariance results is often a matter of the researcher's judgment. The reader should remember that invariance analysis has to do with the replicability, not the interpretaion, of study results. Large effect sizes but poor invariance results will generally indicate that the variables included in the analysis do have a significant effect on the behavior of the study sample but that this effect cannot necessarily be generalized to the larger population.

REFERENCES

Carver, R.C. (1978). The case against statistical significance testing. Harvard Educational Review, 48, 378-399.

Cotton, J.W. (1967). Elementary Statistical Theory for Behavior scientists. Reading, MS: Addison-Wesley.

Fish, L. (1986). Relationships between Gender, Proposed Teaching Level, and Several Background Variables to Motives for Selecting a Career in Teaching. Unpublished master's thesis, Dept. of Educational Leadership & Foundations, University of New Orleans.

Kerlinger, F.N. & Pedhazur, E.J. (1973). Multiple Regression in Behavioral Research. New York, NY: Holt, Rinehart & Winston.

Lykken, D.T. (1968). Statistical significance in psychological research. Psychological Bulletin, 70, 151-159.

Mosier, C.I. (1951). Problems and designs of cross-validation. Educational and Psychological Measurement, 11, 5-11.

Selvin, H.C. (1957). A critique of tests of significance in survey research. American Sociological Review, 22, 519-527.

Thompson, B. (1984). Canonical Correlation Analysis: Uses and Interpretation. Beverly Hills, CA: Sage.

Thompson, B. (April, 1986). A Partial Test Distribution for

Cosines Among Factors Across Samples. Paper presented at the annual meeting of the American Educational Research Association, San Francisco.

Table 1

Hypothetical data set

| | Predictor Variables | | Criterion Variables | |
|---|---|---|---|---|
| | A | B | X | Y |
| Case: | | | | |
| 1 | 0 | 7 | 4 | 7 |
| 2 | 0 | 4 | 2 | 5 |
| 3 | 5 | 3 | 3 | 9 |
| 4 | 2 | 6 | 7 | 2 |
| 5 | 2 | 3 | 0 | 8 |
| 6 | 9 | 9 | 9 | 7 |
| 7 | 5 | 1 | 3 | 0 |
| 8 | 1 | 8 | 8 | 7 |
| 9 | 2 | 2 | 4 | 0 |
| 10 | 5 | 2 | 5 | 8 |

Table 2

Variable values standardized to z-score form within invariance groups

| | Predictor Variables | | Criterion Variables | |
|---|---|---|---|---|
| | A | B | X | Y |
| Group 1: | | | | |
| 1 | -0.878 | 1.321 | 0.309 | 0.288 |
| 2 | -0.878 | -0.330 | -0.464 | -0.432 |
| 3 | 1.562 | -0.881 | -0.077 | 1.009 |
| 4 | 0.098 | 0.771 | 1.468 | -1.514 |
| 5 | 0.098 | -0.881 | -1.236 | 0.649 |
| Group 2: | | | | |
| 6 | 1.470 | 1.216 | 1.236 | 0.644 |
| 7 | 0.192 | -0.899 | -1.082 | -1.089 |
| 8 | -1.086 | 0.952 | 0.850 | 0.644 |
| 9 | -0.767 | -0.635 | -0.696 | -1.090 |
| 10 | 0.192 | -0.635 | -0.309 | 0.892 |

Table 3

Standardized canonical function coefficients

| | | Function 1 | Function 2 |
|---|---|---|---|
| Group 1: | A | 0.971 | 0.724 |
| | B | 1.146 | -0.393 |
| | X | 1.373 | 0.302 |
| | Y | 0.751 | 1.189 |
| Group 2: | A | 0.042 | 1.028 |
| | B | 0.989 | -0.284 |
| | X | 1.174 | -0.958 |
| | Y | -0.249 | 1.494 |

Table 4

Actual and invariance composite scores

| | Actual comp. score | | Invariance comp. score | |
|---|---|---|---|---|
| | Predictor | Composite | Predcitor | Composite |
| Group 1 | | | | |
| 1 | 0.661 | 0.641 | 1.269 | 0.291 |
| 2 | -1.231 | -0.961 | -0.364 | -0.436 |
| 3 | 0.507 | 0.651 | -0.805 | -0.342 |
| 4 | 0.977 | 0.880 | 0.766 | 2.100 |
| 5 | -0.914 | -1.211 | -0.867 | -1.613 |
| Group 2 | | | | |
| 6 | 1.265 | 1.291 | 2.819 | 2.182 |
| 7 | -0.881 | -0.998 | -0.844 | -2.304 |
| 8 | 0.895 | 0.837 | 0.036 | 1.651 |
| 9 | -0.660 | -0.545 | -1.471 | -1.774 |
| 10 | -0.620 | -0.585 | -0.541 | 0.245 |

Table 5

Sqaures of actual and invariance canonical correlation coefficients

| | | Source of function coefficients | |
|---|---|---|---|
| | | Group 1 | Group 2 |
| Source of data | Group 1 | 0.952 | 0.526 |
| | Group 2 | 0.635 | 0.992 |